PROPOSAL FOR TOPIC APPROVAL TO BE SUBMITTED TO THE UNIVERSITY OF MUMBAI FOR THE DEGREE OF **Ph.D. (SCIENCE AND TECHNOLOGY)** IN THE SUBJECT OF **INFORMATION TECHNOLOGY**

Title of Topic	:	Design and Implementation of Teaching Avatars in Virtual Learning Environment: A Case Study Approach
Name of the Student	:	Ms. Jignasha Arjunbhai Sosa
Basic Qualification	:	M.Tech(Computer Engineering) B.E.(Computer Engineering)
In the Case of Ph.D. student Year of Passing of PET/NET/SET/GATE (Teaching Experience)	:	PET - 2022
Name of the Research Supervisor/ Guide	:	Dr. Bhavesh Valji Patel
Date of Superannuation / Retirement	:	20/03/2036
Name of the Research Center	:	Shah and Anchor Kutchhi Engineering College
Signature of Student	:	
Signature of the Research Supervisor	:	

Seal of College

Head / Director / Principal
Research Centre

PRINCIPAL
Shah & Anchor Kutchhi Engineering College
Autonomous Institute
Mahavir Education Trust Chowk,
W. T. Patil Marg, Next to D-Mart,
Chembur - 400088

Sr. no.	Content	Page no.
1	Introduction	1
2	Aims & Objectives	1
3	Literature Review and its Limitations	2
4	Research Hypotheses	4
5	Research Methodology	4
6	Statistical Plan	6
7	Expected Outcome	8
8	References	9

Statement by the Candidate

I wish to state that the work embodied in this synopsis titled "Design and Implementation of Teaching Avatars in Virtual Learning Environment: A Case Study Approach" forms my own contribution to the work carried out under the guidance of Dr. Bhavesh Valji Patel at the Shah and Anchor Kutchhi Engineering College. I declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission.

Name: Ms. Jignasha Arjunbhai Sosa

Roll Number: 124PHIT1001

1. Introduction

The idea of virtual human begins in the year 1995 according to the article published in the journal title "IEEE Transactions on Systems, Man, and Cybernetics". The rapid growth of artificial intelligence is changing many areas of our lives, bringing exciting advancements in virtual human technology. These technologies aim to create systems that can interact with people in realistic, emotional, and meaningful ways. Whether in education, healthcare, entertainment, or customer service, virtual humans are being designed to simulate human-like behavior and communication. However, creating such systems comes with challenges that require careful attention.

Central to these advancements is the pursuit of realism and coherence in virtual interactions. From the synchronization of audio-visual components to the seamless transition of gestures and facial expressions, achieving authenticity remains a key focus. At the same time, creating systems that are adaptable across diverse scenarios and accessible in real-time poses significant technical and computational challenges. Furthermore, reliance on high-quality inputs such as audio and image data underscores the vulnerabilities of these systems when confronted with noisy or unpredictable environments. Despite their potential, virtual humans face some important limitations. For example, there are ethical concerns about their use, such as the risk of bias or emotional manipulation, and issues with protecting user privacy. Additionally, the high cost of running these systems often limits their use to organizations with significant resources. This makes it harder for them to reach wider audiences or adapt to more diverse situations.

This report delves into the evolving landscape of virtual human technologies, exploring their foundational principles, applications, and inherent challenges. By critically examining both opportunities and constraints, the discussion aims to illuminate pathways for future research, ensuring that innovation in this domain is guided by technical excellence, ethical foresight, and practical viability.

2. Aims & Objective

The aim is to create an intelligent virtual human interface that leverages Large Language Models for seamless audio and video interaction. The objectives are as follows,

- Develop Virtual Human Avatar with LLM with high output accuracy.
- Develop Virtual Human Avatar with audio.
- Develop Virtual Human Avatar must operate for maximum 30 minutes based on user input.

3. Literature Review and its Limitations

This section describes about existing systems and its limitations.

The authors has described [1], a method for generating virtual human videos by employing constrained keyframes to maintain consistency in character features across frames. A single-mode transformer encoder processed both visual and audio inputs, synchronizing facial expressions with audio for seamless transitions. The authors has highlighted the use of a silent state to connect video segments. Limitations included reduced control over intermediate frames, reliance on high-quality input, computational costs, and challenges in generalizing to diverse scenarios. The authors has suggested more detailed evaluation metrics were needed to validate performance.

The authors has explained [2], the potential of ChatGPT as a virtual teaching assistant, highlighting its strengths in engagement, coherent communication, and personalized guidance. However, they detected ideological biases in ChatGPT's responses due to its training data, raising ethical concerns about its impact on learners. The authors has used Appraisal Theory and Systemic Functional Linguistics to analyze the language patterns, revealing significant ethical challenges. The limitations were acknowledged in the scope of their analysis, the evolving nature of LLMs, and the lack of comparisons with human tutors.

The authors has provided [3], a survey of interaction design for 3D virtual humans, categorizing modalities such as speech, eye movement, posture, and facial expressions. The importance of multimodal integration were emphasized for creating realistic and engaging virtual humans. The authors has discussed applications in education, healthcare, and workplaces, while identifying challenges such as emotion recognition and low-cost development. However, the survey focused on high-level overviews rather than detailed technical discussions or empirical evaluations.

The authors introduced [4], a virtual human platform driven by large language models (LLMs) to facilitate realistic, open-ended conversations for eliciting emotions. The integration of multimodal inputs like realistic avatars, facial expressions, and voice synthesis to enhance user interaction were highlighted. The authors has also described a modular framework incorporating psychological constructs such as personality and mood. The limitations such as constrained emotion elicitation, VR platform constraints, and the lack of long-term interaction evaluations were acknowledged. Ethical concerns, particularly in sensitive applications, were also discussed.

The authors presented [5], Persona-Gestor, a model for synthesizing personalized 3D gestures from speech using a fuzzy feature extraction strategy. The authors has also described the use of a non-autoregressive Adaptive Layer Normalization Transformer Diffusion architecture to generate high-quality gestures efficiently. The authors has reported superior performance compared to state-of-the-art models but noted challenges such as computational costs, limited generalization to unseen speakers, and dependence on large datasets. The suggestions about emotional expressiveness and incorporating conversational context for future improvements were expanded.

The authors has proposed [6], a method for generating high-quality 3D human textures using weak supervision, eliminating the need for ground-truth 3D texture data. The authors has explained the combination of texture transfer and UV-map-based regression, improving texture quality for asymmetric structures. Applications included person re-identification tasks, where augmented datasets were

created. The authors has also discussed limitations such as segmentation dependency, computational costs, and difficulties in handling diverse clothing styles or complex poses.

The authors has described [7], Audio2AB, a method for generating synchronized animations of facial expressions and body gestures using audio inputs. The the use of parameterized animations and a novel GAN-GF loss function for training stabilization were highlighted. The authors has also emphasized the multimodal nature of the system, which incorporated text, emotion, and semantic relevance labels. However, the challenges of labeled data dependency, limited generalization to diverse character models, and the complexity of achieving real-time performance were acknowledged.

The authors has provided [8], a detailed survey of virtual human technologies, covering their history, development, and core components such as motion capture, speech synthesis, and artificial intelligence. The diverse applications in education, entertainment, and customer service while discussing challenges like computational costs and the uncanny valley effect were explored. Future trends such as deep learning advancements and virtual human integration into the metaverse were also addressed. The authors has acknowledged the breadth of the survey limited in-depth analysis of specific techniques.

The authors has proposed [9], a two-stage pipeline to create virtual humans for customer service, focusing on action video generation and emotional lip synchronization. The the creation of a novel dataset of human videos capturing service actions and emotional talking faces were emphasized. While the method improved user engagement, the authors has also acknowledged limitations in action diversity, emotional expressiveness, and adaptability to diverse scenarios. Comprehensive evaluation metrics were suggested for better assessment.

The authors has introduced [10], MSStyleTTS, a model for expressive speech synthesis using multi-scale style modeling at subword, sentence, and global levels. The authors has also explained the hierarchical context encoding mechanism that improved coherence across sentences. By employing residual connections and autoregressive predictors, the authors also has achieved improved expressiveness in speech synthesis. However, they noted challenges such as data dependency, computational costs, and limited generalization to unseen speaking styles.

The authors has investigated [11], how mismatches between photorealistic virtual human appearance and synthetic voice realism affected user perception. The authors has also described a large-scale VR experiment that examined social presence, empathy, and emotional responses to virtual humans. The trade-offs between expressiveness and naturalness in synthetic voices, emphasizing the importance of voice consistency were discussed. The limitations such as short-term interaction focus and the need for more naturalistic, long-term studies were acknowledged.

4. Research Hypotheses

There are four hypotheses involved in these system which are provided as follows,

- Hypotheses 1: Can audio be embedded to human avatar.
- Hypotheses 2: Can human avatar provide efficient output than ChatGPT.
- Hypotheses 3: Can facial expression and lip-sync provide expected output.
- Hypotheses 4: Can human avatar be operated as teaching assistant.

5. Research Methodology

The methodology specifies a systematic approach to achieve the research objectives. It involves research design, research framework, data collection, data analysis, tools and technologies and ethical considerations. The below figure represents research plan for the proposed work.

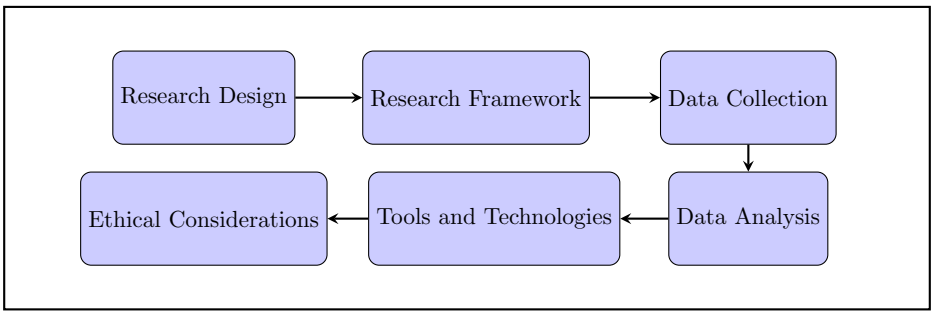

Fig. 5.1 Research Plan

The research plan phases are discussed in below sections.

Phase 1: Research Design

The proposed work will be based on a hybrid research design to achieve a balance between experimentation, user experience, and system performance. The experimental approach will be applied to develop and test a system for creating human avatars that accurately combine audio and video. To gain insights into the quality of the system, qualitative methods will be used to explore how users perceive the realism and ease of interaction with the avatars. At the same time, quantitative methods will focus on measuring technical performance aspects, such as how well the system synchronizes audio with video, the consistency of the visual elements, and the speed of real-time rendering.

Phase 2: Research Framework

The research framework is structured into several interconnected phases. First, a literature review will be conducted to study the role of large language models in multimedia applications and the techniques used for synthesizing audio and video content. Following this, a system architecture will be developed that integrates large language models for natural language processing, audio tools for extracting features like pitch and tone, and advanced video modeling systems for realistic avatar rendering. A multimodal fusion mechanism will ensure seamless synchronization of audio and video. The proposed framework will introduce keyframe constraints using anchor frames and interpolation

techniques to maintain smooth and consistent video transitions. Finally, a working prototype of the system will be created using tools like Python, machine learning libraries, and rendering software.

Phase 3: Data Collection

Data collection will focus on gathering high-quality resources for both audio and video. Speech datasets will be required to provide the audio content for the avatars, ensuring diverse and realistic speech inputs. Similarly, facial animation datasets will be collected to create video content, with an emphasis on ensuring these datasets are varied enough to train the system for a wide range of scenarios. These datasets will be carefully selected to align with the system's goal of achieving accurate and lifelike avatar representation.

Phase 4: Data Analysis

The data analysis process will evaluate the system's performance through both numerical metrics and user feedback. Quantitative analysis will measure factors like how accurately the avatar's lip movements match the audio, the smoothness of transitions between video frames, and the overall processing speed of the system during real-time rendering. A user survey will be conducted to collect qualitative feedback on how realistic the avatars appear, how easy they are to interact with, and their potential applications in settings like virtual assistants or online learning platforms. A comparative analysis will be performed to benchmark the system against existing solutions, highlighting its advantages and areas for improvement.

Phase 5: Tools and Technologies

The implementation of the system will rely on advanced tools and technologies to ensure high-quality results. Large Language Models, such as GPT, can be used to understand and generate natural language. For audio processing, tools like Wav2Vec extract essential features, while video generation rely on advanced rendering technologies like DALL-E, StyleGAN3, and Neural Radiance Fields. Development and testing will be carried out using programming languages and frameworks like Python, PyTorch, and TensorFlow, while rendering software such as Unity and Blender may be used to create and fine-tune the visual representation of the avatars.

Phase 6: Ethical Considerations

Ethical considerations are a core component of this study to ensure responsible and unbiased results. It ensures whether the datasets comply with privacy regulations, and the explicit consent will be obtained for any data involving human participants. It will also avoid biases in the generation of avatars, ensuring that they are culturally and gender-sensitive. By following these practices, the study will aims to create a fair and inclusive system while adhering to the highest ethical standards.

6. Statistical Plan

The below table contains the statistical plan for the research. The table 6.1 describes the progress plan for time period January to June 2025.

Table 6.1 Time Period Months 1–6 (January to June)

Sr. no.	Goals and Activities	Deliverables/ Outcomes
1	Literature Review: Study existing works on LLMs, audio-visual synthesis, and human avatar technology.	Comprehensive understanding of research gaps. Draft research questions and hypotheses.
2	Define key hypotheses (e.g., LLM-based avatars improve realism compared to existing methods).	Null and alternative hypotheses. Conceptual research framework.
3	Data Collection Planning: Identify datasets for training and testing.	Finalized datasets for audio, video, and multimodal training/testing.
4	System Architecture Design: Develop a blueprint for integrating LLMs with audio and video models.	Initial design for the avatar system architecture.
5	Research Publication: Document results and submit papers to journals/conferences.	Research articles and presentations highlighting contributions.

The table 6.2 describes the progress plan for time period July to December 2025.

Table 6.2 Time Period Months 7–12 (July to December)

Sr. no.	Goals and Activities	Deliverables/ Outcomes
1	Prototype Development: Build the first version of the audio-visual representation system using existing datasets.	Functional prototype capable of generating basic avatars from text inputs.
2	Algorithm Implementation: Implement keyframe constraints and multimodal fusion.	Enhanced system for stable and synchronized audio-video generation.
3	Testing: Test the prototype on a small dataset to assess performance.	Performance metrics for lip-sync accuracy, frame consistency, and realism.
4	Feedback Integration: Refine system based on tested results and feedback from early users.	Improved prototype with better realism and stability.
5	Research Publication: Document results and submit papers to journals/conferences.	Research articles and presentations highlighting contributions.

The table 6.3 describes the progress plan for time period January to June 2026.

Table 6.3 Time Period Months 13–18 (January to June)

Sr. no.	Goals and Activities	Deliverables/ Outcomes
1	Extended Testing: Conduct large-scale tests using large datasets.	Performance evaluation on varied accents, languages, and facial features.
2	Statistical Analysis: Test hypotheses using collected data.	Statistical report validating or rejecting hypotheses.
3	Optimization: Enhance processing speed and reduce latency in avatar generation.	Optimized avatar system capable of near-real-time operation.
4	Validation: Use cross-validation techniques to ensure consistency and reliability of results.	Validated model performance metrics with reproducible results.
5	Research Publication: Document results and submit papers to journals/conferences.	Research articles and presentations highlighting contributions.

The table 6.4 describes the progress plan for time period July to December 2026.

Table 6.4 Time Period Months 19–24 (July to December)

Sr. no.	Goals and Activities	Deliverables/ Outcomes
1	User Study: Conduct qualitative user surveys and experiments to assess satisfaction, interaction quality, and realism.	User feedback on the effectiveness of the avatar system.
2	Comparative Study: Compare the proposed system with existing technologies (e.g., baseline models).	Benchmarking report comparing the new system with existing methods.
3	Final System Deployment: Integrate findings into the final version of the avatar system.	Fully developed avatar system ready for practical applications.
4	Final document/ report/ thesis preparation: Organize research into a structured, well-structured document for submission.	A well-structured thesis document ready for submission, including research findings, methodology, analysis, discussion, and references.

7. Expected Outcomes

The proposed system will be able to,

- **Generate Synthetic Human Avatar:** This outcome involves the creation of a digital human persona through sophisticated algorithms and 3D modeling techniques. This includes the accurate rendering of human features, as well as the simulation of lifelike movements and behaviors that replicate those of real humans, achieving a high level of realism in the virtual representation.
- **Phonetic Synthesis and Affective Computing Integration:** This outcome encompasses the incorporation of advanced speech synthesis technologies to enable the virtual human to engage in vocal communication. Additionally, affective computing techniques are employed to imbue the virtual persona with the capability to exhibit emotions through vocal intonation and facial expressions, enhancing the naturalness and emotional depth of interactions.
- **Large Language Model Data Assimilation and Response Generation:** This outcome involves the extraction of relevant data from extensive language models and its seamless integration into the virtual human's cognitive architecture. This enables the digital entity to generate contextually appropriate and coherent responses, with the added capability of vocal expression enhanced by emotional cues and accompanying gestural animations, thereby delivering a comprehensive and immersive user experience.

References

[1] P. Li, X. Li and L. Yin, "Research and Implementation of Constrained Keyframe-Based Virtual Human Video Generation," in IEEE Access, vol. 12, pp. 137089-137104, 2024.

[2] Margo Van Poucke, "ChatGPT, the perfect virtual teaching assistant? Ideological bias in learner-chatbot interactions," in Elsevier Computers and Composition Elsevier Journal, vol. 73, p. 102871, 2024.

[3] X. Wang, N. Cao, Q. Chen, S. Cao, "The interaction design of 3D virtual humans: A survey," in Computer Science Review Elsevier Journal, vol. 53, p. 100653, 2024.

[4] J. Jurado, L. Zaragozá, M. Minissi, M. Alcañiz, J. Morales, "Developing conversational Virtual Humans for social emotion elicitation based on large language models," in Expert Systems with Applications Elsevier Journal, vol. 246, p. 123261, 2024.

[5] F. Zhang, Z. Wang, X. Lyu et al., "Speech-Driven Personalized Gesture Synthetics: Harnessing Automatic Fuzzy Feature Inference," in IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 10, pp. 6984-6996, Oct. 2024.

[6] F. Zhao, S. Liao, J. Huo, W. Wang et al., "Weakly Supervised Joint Transfer and Regression of Textures for 3-D Human Reconstruction," in IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 4400-4410, Feb. 2024.

[7] L. Niu, W. Xie, D. Wang, Z. Cao, X. Liu, "Audio2AB: Audio-driven collaborative generation of virtual character animation," in Virtual Reality Intelligent Hardware Elsevier Journal, vol. 6, no. 1, pp. 56-70, 2024.

[8] L. Cui and J. Liu, "Virtual Human: A Comprehensive Survey on Academic and Applications," in IEEE Access, vol. 11, pp. 123830-123845, 2023.

[9] Y. Sun, Z. Sun, Y. Wen, S. Ye, T. Lv, M. Yu, et al., "Generation of virtual digital human for customer service industry," in Computers Graphics Elsevier Journal, vol. 115, pp. 359-370, 2023.

[10] S. Lei et al., "MSStyleTTS: Multi-Scale Style Modeling With Hierarchical Context Information for Expressive Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3290-3303, 2023.

[11] D. Higgins, K. Zibrek, J. Cabral, D. Egan, R. McDonnell, "Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans," in Computers Graphics Elsevier Journal, vol. 104, pp. 116-128, 2022.

[12] J. Yu and Z. -F. Wang, "A Video, Text, and Speech-Driven Realistic 3-D Virtual Head for Human–Machine Interface," in IEEE Transactions on Cybernetics, vol. 45, no. 5, pp. 991-1002, May 2015.

[13] H. Tang, Y. Fu, J. Tu, M. Hasegawa-Johnson and T. S. Huang, "Humanoid Audio–Visual Avatar With Emotive Text-to-Speech Synthesis," in IEEE Transactions on Multimedia, vol. 10, no. 6, pp. 969-981, Oct. 2008.

[14] Y. Fu, R. Li, T. S. Huang and M. Danielsen, "Real-Time Multimodal Human–Avatar Interaction," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 4, pp. 467-477, April 2008.

[15] Y. Fu and N. Zheng, "M-Face: An Appearance-Based Photorealistic Model for Multiple Facial Attributes Rendering," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 7, pp. 830-842, July 2006.

[16] N. M. Thalmann, P. Kalra and M. Escher, "Face to virtual face," in Proceedings of the IEEE, vol. 86, no. 5, pp. 870-883, May 1998.

[17] P. M. Jones and C. M. Mitchell, "Human-computer cooperative problem solving: theory, design, and evaluation of an intelligent associate system," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 25, no. 7, pp. 1039-1053, July 1995.

[18] Z. Ying, "Experience of intelligent speech robot in music online classroom based on deep learning and virtual reality," in Entertainment Computing Elsevier Journal, vol. 52, p. 100795, 2025.